May 19, 2026

ISSUE BRIEF | AI and Emerging Technology

Accelerating AI Agent Adoption Through Improved Security

Jack Crovitz, Cole Salvador, & Yusuf Mahmood

TOPLINE POINTS

- Rapid adoption of AI agents can make American enterprises more productive, federal agencies more efficient, and American warfighters more lethal.
- Many institutions, especially in the U.S. government, are slow to deploy this technology because of legitimate security concerns. These concerns include AI agents' tendency to behave in unpredictable ways and their susceptibility to adversarial attacks such as model poisoning and agent hijacking.
- The U.S. government can accelerate the adoption of AI agents by promoting strong security standards for AI development, publishing clear guidance on the secure deployment of AI agents, and investing in technical research in AI reliability.

Introduction

Agentic artificial intelligence (AI) systems are poised to transform our economy and national security. Rather than simply responding to prompts, AI agents can plan multi-step projects, take autonomous actions, and iterate toward goals with minimal human intervention. Unlike traditional AI applications such as chatbots, agents can access and use tools that enable interaction with the digital or physical world. These tools often include real-time code execution, web browsing, file management, and open-ended computer use (RFI Regarding Security Considerations for AI Agents, 2026).

AI agent technology holds enormous promise for American enterprises, federal agencies, and warfighters. In the private sector, AI agents are already empowering American workers to focus on creative, strategic, and high-value work rather than tedious drudgery. For federal agencies, they can allow faster, more accurate processing of the complex regulatory and administrative burdens that slow mission execution. As the White House's *AI Action Plan* states: "With AI tools in use, the Federal government can serve the public with far greater efficiency and effectiveness" (Kratsios et al., 2025). The Department of Energy (DOE) already plans to deploy AI agents to automate scientific research via the Genesis Mission (Trump, 2025b; Kohli & Lue, 2025). AI agents will also help our warfighters through

real-time decision support, autonomous reconnaissance, and logistics optimization—in situations where speed and precision are matters of life and death.

Unfortunately, many enterprises and government agencies are delaying adoption because the same qualities that make AI agents useful—autonomy and tool use—also make them hard to control. Government agencies recognize that AI agents raise security issues fundamentally distinct from those associated with chatbots or conventional software (RFI Regarding Security Considerations for AI Agents, 2026). According to the National Institute of Standards and Technology's (NIST) Center for AI Standards and Innovation (CAISI), such concerns about AI agent security are currently impeding adoption within U.S. government agencies (NIST, 2026). The Department of Labor's chief AI officer, for instance, has observed that the distinctive capabilities of AI agents demand stronger safety and security measures than those applied to other AI tools (Heckman, 2026).

AI agent security issues are also causing harm in the private sector. In July 2025, for example, an AI agent owned by the startup Replit deleted an entire live production database, despite explicit user instructions not to do so (Nolan, 2025). More recently, an agent managed by the Director of Alignment at Meta Superintelligence, Summer Yue, began to "speedrun" deleting Yue's entire email inbox—even after many attempts to stop the agent (Cramer, 2026). According to Yue's account, she had to physically run to her desktop to shut the AI agent down. If the head of alignment at one of the world's leading AI companies cannot control her own AI agent, the security challenges for private enterprises and government agencies are clear. These examples show why anomalous outputs are much more dangerous for autonomous AI agents than for many traditional AI systems. Chatbots cannot delete your inbox, sabotage your deployed codebase, or leak sensitive information. AI agents can.

These security challenges have prevented America's companies and government from harnessing the full potential of agents. Many organizations cite the lack of strong, reliable, and effective AI agent security standards as a reason to delay or abandon AI agent deployment initiatives (Antone & Chang, 2025). The challenge is most dire in U.S. government agencies, many of which could derive enormous value from AI agents but fail to deploy the technology due to a risk-averse culture and fears about security vulnerabilities. NIST's CAISI states that AI agent security concerns in U.S. government agencies "hinder adoption today" (NIST, 2026).

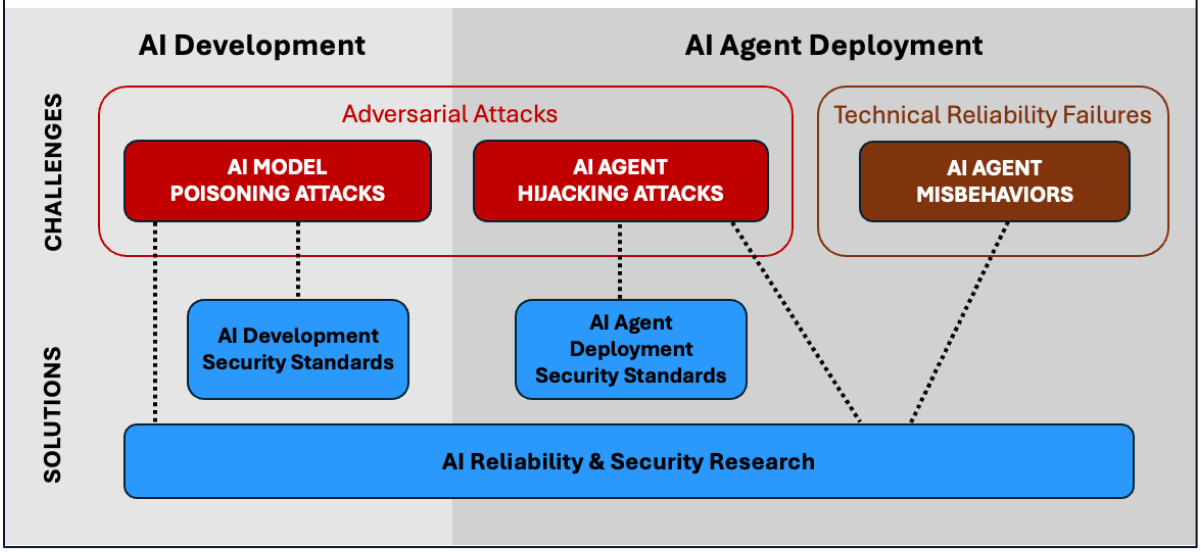
In this Issue Brief, we examine three major challenges to AI agent security:

1. **AI Model Poisoning**, a type of adversarial attack that sabotages or compromises an AI model during the training process;
2. **AI Agent Hijacking**, a type of adversarial attack that allows malicious actors to gain control of AI agents during deployment; and
3. **AI Agent Misbehaviors**, which occur because developers lack the technical methods to guarantee reliable behavior.

Public-private sector collaboration can address all three of these concerns. We suggest three broad strategies for federal policymakers to break the security bottleneck on AI agent adoption for all prospective users in both the private and government sectors:

1. **Establishing AI development security standards** for frontier model companies;
2. **Promoting agent deployment security standards** for institutions that use AI agents.
3. **Supporting technical research** in methods to ensure AI reliability and security.

Figure 1: *Outline of the Major Challenges and Solutions in AI Agent Security*



These strategies will accelerate the adoption of cutting-edge AI across the nation, advancing President Trump's vision of "a new Golden Age of innovation, human flourishing, and technological achievement for the American people" (Trump, 2025a).

Risk: AI Model Poisoning Attacks

Model poisoning is a type of adversarial attack in which bad actors corrupt an AI model by contaminating its training data. Because frontier models are trained on trillions of tokens—in large part scraped from the open internet—the potential for exploitation is enormous (Villalobos et al., 2022). One well-documented method involves inserting a small number of manipulated documents into a model's training corpus to implant a "backdoor": a hidden trigger that causes the model to behave dangerously under specific conditions while appearing normal otherwise (Vassilev et al., 2025; Banerjee & Aarne, 2026).

Recent research suggests that these attacks may be both surprisingly easy to execute and exceptionally difficult to detect. In October 2025, researchers demonstrated that as few as 250 malicious documents could successfully implant backdoors into models with 13 billion parameters (Souly et al., 2025). Alarmingly, the data required to poison a model does not appear to scale with model size—meaning a sophisticated actor could compromise one of today's most capable systems by tampering with only a tiny fraction of its training data. After they have been inserted, such backdoors are nearly impossible to detect and, even if discovered, difficult to remove (Hubinger et al., 2024). In some cases, anti-backdoor training seemingly teaches the AI model to conceal its backdoor rather than eliminate it.

Data poisoning is not a theoretical concern. In 2025, a lone Western security researcher cheaply installed a backdoor in the Chinese near-frontier AI model DeepSeek-R1. They did so by manipulating text in public code repositories that DeepSeek later scraped to train R1 (Banerjee & Aarne, 2026). Once R1 was released to the public, the researcher used the artificial backdoor to violate the model's guardrails. This real-world case was inspired by prior experiments, which demonstrated that production models can be sabotaged for less than $100 by changing data on the open internet (Carlini et al., 2024; Souly et al., 2025).

If an individual can successfully backdoor China's most advanced model, American AI companies are certainly vulnerable to far more serious attacks from nation-state actors. The capabilities that Chinese state-sponsored hackers could bring to bear on a deliberate model poisoning campaign are formidable (Banerjee & Aarne, 2026). Such actors could use compromised insiders to undermine or exfiltrate data filtering infrastructure; buy domains used as training data sources to quietly swap in malicious content; or simply flood the pipeline with poisoning attempts at such a scale that even marginal success rates would corrupt the resulting AI model's behavior.

Indeed, state-backed hackers could attempt far more subtle and destructive model poisoning methods than simple backdoor insertion. Researchers theorize that attackers could use model poisoning to inject "sophisticated secret loyalties" into a model that would make it autonomously work on behalf of that attacker (Banerjee & Aarne, 2026). Such a model might, for example, insert subtle software vulnerabilities into critical infrastructure and feign incompetence on tasks that would promote American technological advancement. American AI developers would be unwittingly training "sleeper agents" for the Chinese Communist Party (Miyazono, 2025).

Such attacks could have catastrophic consequences in deployment. Software developers already use AI systems to code; corrupted AI models could insert security vulnerabilities into critical infrastructure or U.S. government systems (Gent, 2025). AI agents are also being used in military environments, including targeting systems, where corrupted models could cause substantial damage to

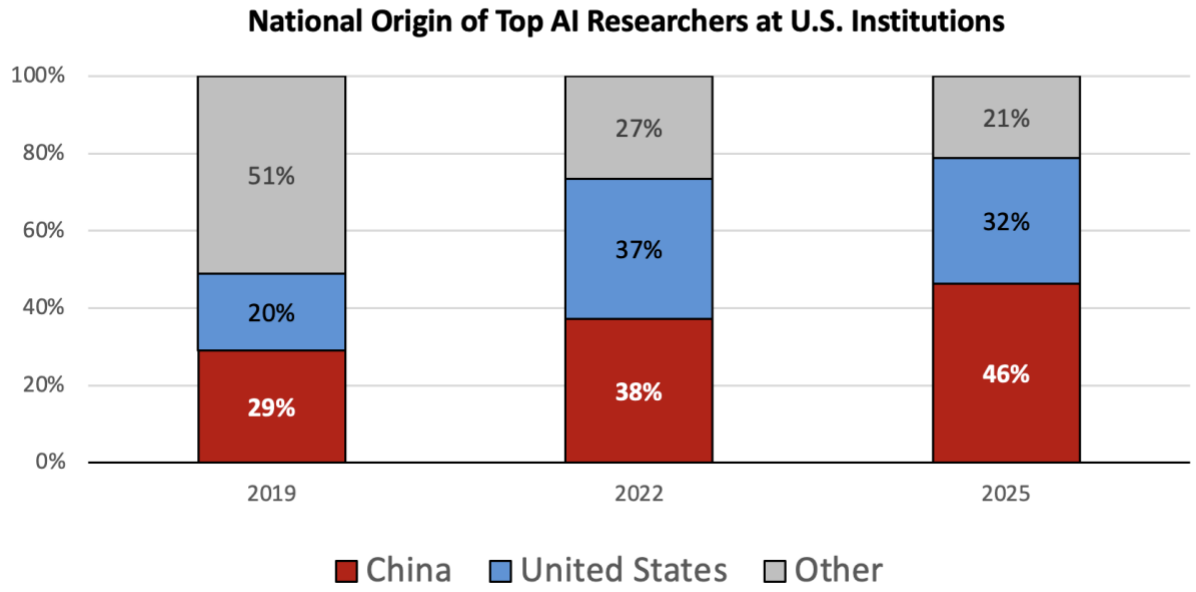
national security (Jensen & Strohmeyer, 2025). As AI agents become more integrated in American critical infrastructure, the Pentagon, and the intelligence community, we must assume that adversaries will try to poison the AI models (if they have not done so already). The evidence suggests that they will succeed.

Security Vulnerabilities That Make American AI Developers Prone to Poisoning Attacks

There are five main categories of vulnerabilities that malicious actors could target to compromise the integrity of an AI company's model development supply chain and successfully poison their deployed AI models: (i) compromised insider personnel, (ii) cybersecurity vulnerabilities, (iii) data center insecurity, (iv) flawed data hygiene protocols, and (v) insecure model auditing protocols.

Compromised insider personnel represent a serious security threat to frontier AI developers. American AI companies are known to employ many foreign nationals as researchers and engineers, particularly Chinese: more than 40% of the top AI researchers at American AI companies and research institutions received their undergraduate education in China (Paulson Institute, 2025; see also Sheehan & Zhuang, 2025). Article 7 of China's National Intelligence Law requires all citizens to cooperate with Beijing's intelligence network, leaving these researchers little choice but to comply with Chinese demands (Doshi, 2024). In the context of Beijing's aim to steal, subvert, and sabotage American AI development by whatever means necessary, this represents a direct security threat to the integrity of the AI supply chain (Mahmood et al., 2026).

Figure 2: *Individuals from China now outnumber Americans among top technical researchers at American AI companies and research institutions. China's National Intelligence Law requires all citizens to cooperate with Beijing's intelligence network.*



Note. The methodology behind this data uses the location of a researcher's undergraduate education as a proxy for national origin, so it likely underestimates the share of Chinese-origin researchers. Data from the *Global AI Talent Tracker*, by the Paulson Institute, 2020, 2022, 2025 (retrieved from https://www.paulsoninstitute.org/press_release/the-paulson-institutes-think-tank-macropolo-releases-global-ai-talent-tracker, https://www.edwardconard.com/macro-roundup/57-of-the-top-

2-of-global-ai-talent-currently-works-in-the-united-states-of-the-top-20-of-ai-workers-in-the-us-37-are-from-the-us-and-38-are-from-china-macropolochina, and https://archivemacropolo.org/interactive/digital-projects/the-global-ai-talent-tracker) and author's calculations.

Analysts believe that all of the frontier AI companies are penetrated by foreign intelligence agents, particularly from China (Harris & Harris, 2025). One Chinese AI researcher at Google was convicted in 2026 of exfiltrating proprietary American AI technology to China (Department of Justice, 2026). Compromised insiders could facilitate AI model poisoning by sabotaging data hygiene systems, directly inserting manipulated data into the training process, and subverting model auditing systems. The attack surface is vast, and American AI developers are not ready.

Cybersecurity vulnerabilities are a major weakness at many large American AI developers. Former insiders report that the companies "treat security as an afterthought," and cybersecurity professionals acknowledge that these organizations are not prepared to resist sophisticated cyberattacks by nation-state adversaries (Aschenbrenner, 2024; Harris & Harris, 2025). Unfortunately, security staff at AI companies report being discouraged from speaking publicly about major security vulnerabilities, so it is difficult to understand the magnitude of these vulnerabilities. At least one high-profile firing—that of former OpenAI researcher Leopold Aschenbrenner—has been linked to raising concerns about company security (Altchek, 2024). We know that nation-state adversaries have already begun attempting cyberattacks against large American AI developers. OpenAI reports that a Chinese cyber-espionage group called "SweetSpecter" has been seeking to infiltrate the company through spear phishing, a form of personalized hacking (OpenAI, 2024).

Data center insecurity is a widely acknowledged challenge in the frontier AI development community. The physical infrastructure required to train large AI systems—including servers, cooling systems, and networking hardware, concentrated in "training clusters"—represents a significant attack surface that extends beyond digital intrusions. Experts warn that data centers are vulnerable not only to cyberattacks but also to compromised personnel and physical sabotage (Nevo et al., 2024). Infiltration of training clusters could allow malicious actors to "interfere with AI model training to later compromise the model's users at scale" by poisoning the AI model being trained (Grunewald & Gershovich, 2025).

Flawed data hygiene protocols would allow maliciously manipulated data to enter the training process for a frontier AI model. "Data hygiene" is the use of filtering, attribution, and automated detection to prevent models from being trained on poisonous data. Because training data sets are massive, and in large part scraped from the internet, data filtering techniques must have a minuscule false-negative removal rate to be effective (Villalobos et al., 2022). Developers must also be careful to track the provenance of training data to detect attempts by malicious actors to smuggle manipulated data into the training set. Researchers believe that frontier AI companies' existing data hygiene methods are far from sufficient, and could likely be fooled by sophisticated and well-resourced adversaries (Draganov et al., 2026). DeepSeek's data hygiene protocols, for example, were unable to detect the manipulated text in certain code repositories that were used to poison their near-frontier model R1 (Banerjee & Aarne, 2026). American AI companies may not be far ahead of their Chinese competitors.

Insecure model auditing protocols may be the most concerning vulnerability in the AI development supply chain. Before releasing a new AI model, frontier AI companies audit the model for backdoors, hidden loyalties, strange behavior, and other warning signs of tampering (Marks et al., 2025). Unfortunately, current methods of uncovering tampering are unreliable, and detecting model poisoning is generally considered an "open challenge" (Chhabra et al., 2026). Some nascent methods, like automated "alignment audits," do show promise of eventually being able to

uncover tampering (Bricken et al., 2025). However, it is unclear whether current auditing methods are sufficient to consistently detect AI model poisoning, and the challenge is heightened because auditing systems may themselves be vulnerable to tampering from cyberattacks or compromised insiders.

Solution: Promoting AI Development Security Standards

The U.S. government can effectively encourage higher security standards in the AI development process at frontier labs. We know that concerns about AI model poisoning are actively impeding the deployment of AI agents in enterprises and government agencies, particularly the Department of War (CNBC, 2026). The policy recommendations below can assuage these concerns and allow our warfighters to deploy the most effective and lethal tools.

Solution 1: The Department of Justice (DOJ) and the Federal Trade Commission (FTC) could issue guidance clarifying that AI developers may share threat intelligence and coordinate security protocols without violating federal antitrust law. Currently, frontier AI labs are deterred from voluntarily collaborating on security research and standards due to the threat of antitrust scrutiny (Michaels et al., 2026; Reinauer, 2026). This fear creates a mismatch of incentives, because it is in the American national interest that its AI companies freely share threat intelligence and voluntarily coordinate advanced security standards. Federal antitrust law is not meant to hamstring critical industries from improving their security. It is certainly not meant to increase the chances of Chinese state-sponsored hackers poisoning American AI models.

The Biden Administration supercharged this dysfunction in December 2024 by withdrawing the FTC's "Antitrust Guidelines for Collaborations among Competitors" (Felstead, 2026). The FTC and the DOJ could draft new guidelines that reinstate the "Antitrust Guidelines for Collaborations among Competitors" and expressly affirm that security coordination among AI developers does not run afoul of federal antitrust law.

Solution 2: The Department of War could use adherence to anti-poisoning security standards as a procurement condition for AI contracts. Under Secretary of War for Research and Engineering Emil Michael identifies "insider threats" at AI labs and the possibility of "model poisoning" as concrete concerns for AI use in the Department of War (CNBC, 2026). Under Secretary Michael is correct that model poisoning threatens the integrity of AI systems and that compromised AI agents pose unacceptable risks to American warfighters. The first step to solve this challenge must be publishing strong security standards for frontier AI labs and using them as conditions for the Department of War's procurement of AI tools.

The statutory foundation of such procurement conditions already exists. The 2026 National Defense Authorization Act (NDAA) directs the Secretary of War to "develop a framework for the implementation of cybersecurity and physical security standards and best practices relating to covered artificial intelligence and machine learning technologies to mitigate risks" associated with warfighters' use of AI technology (S. 1071, 2025, Sec. 1513). It also directs the Pentagon to issue regulations that ensure that this framework "imposes requirements for security on contractors that are designed to mitigate the cybersecurity risks posed by the cyber threat actors" (S. 1071, 2025, Sec. 1512). The risk of adversarial model poisoning of frontier AI systems is serious enough to necessitate specific security provisions under the 2026 NDAA. Such procurement conditions should include provisions for training data hygiene, cyber infrastructure security, hardware security, personnel security, and model auditing protocols.

Solution 3: Congress could impose minimum security standards on all AI developers regardless of their status as military contractors. While the AI security standards initiative included by the 2026 NDAA is a valuable first step, it has two weaknesses. First, the Department of War is not required to draft or implement the standards on any particular timeline, which does not fit the urgency of the challenge. Second, the standards only apply to AI developers that provide services to the Pentagon—which no longer includes all frontier AI companies (Freifeld & Seetharaman, 2026). These weaknesses can be easily overcome. Congress could strengthen the AI development security initiative by directing the U.S. government to impose minimum security standards on all large AI developers within a few months of the legislation's enactment. These standards should not be limited to cybersecurity, but also include guidance on data center security, personnel security, data hygiene, and pre-release AI model auditing.

Solution 4: Congress could prohibit employment discrimination against whistleblowers reporting AI security vulnerabilities. Despite the national security implications of security vulnerabilities at large AI developers, the security staff at these companies report being discouraged from informing policymakers and the public about major issues (Ghaffary, 2024; Verma et al., 2024). One high-profile firing, that of former OpenAI researcher Leopold Aschenbrenner, was likely due to Aschenbrenner raising concerns about OpenAI's security (Altchek, 2024). Reports also suggest that it is conventional in the AI industry for companies to impose extremely restrictive nondisclosure agreements on researchers that prevent them from criticizing their employer's security practices (Field, 2024). Information about the security situation at these companies should not be systematically withheld from policymakers, and whistleblower protections are a time-tested way to prevent such schemes. As Sen. Grassley has explained: "Today, too many people working in AI feel they're unable to speak up when they see something wrong. Whistleblowers are one of the best ways to ensure Congress keeps pace as the AI industry rapidly develops" (Senate Committee on the Judiciary, 2025).

Solution 5: The Center for AI Standards and Innovation (CAISI) could publish voluntary security frameworks and standards for frontier AI developers. These initiatives would establish baseline industry expectations for data hygiene, personnel security, infrastructure security, and model auditing. These would help AI developers proactively design and implement defenses against model poisoning, protecting both the labs and the users of their models. Guidelines would be most effective if they suggest thresholds at which developers could consider upgrading their model development security practices. An example threshold could be an attempted state-level model poisoning effort. CAISI's existing work on model poisoning and AI agent security provides a strong foundation for these standards (Vassilev et al., 2025; CAISI, 2025a; Hamin & Edelman, 2025). The Department of War and other agencies with high security needs could also draw on these standards for drafting procurement conditions.

Solution 6: The National Security Agency (NSA) and the Cybersecurity and Infrastructure Security Agency (CISA) could lead red-teaming exercises that aim to uncover security vulnerabilities in the AI development supply chain and work with industry to resolve those vulnerabilities. NSA and CISA have deep expertise in offensive and defensive cyber operations, including classified threat vectors, so they are uniquely positioned to simulate the tactics of nation-state adversaries. These red-teaming exercises could target the entire AI development process, from data scraping and filtering systems to model auditing processes. They could cover the full set of attack vectors, including cyberattacks, physical intrusion into data centers, and threats from compromised insiders. NSA and CISA could collaborate with CAISI and other agencies to ensure these exercises comprehensively simulate real-world attacks from nation-state adversaries (CAISI, 2025b). The exercises should be conducted with the voluntary cooperation of frontier AI companies.

Red-teaming programs like this one have strong precedent. In other critical industries, NSA and CISA red-teaming exercises have enabled private actors to implement strong security standards and to continuously identify and patch vulnerabilities (Freedberg, 2023; CISA, 2023). American leadership in AI is a "national security imperative," as President Trump affirms, and NSA red-teaming would reflect that strategic significance (Kratsios et al., 2025).

Risk: AI Agent Hijacking Attacks

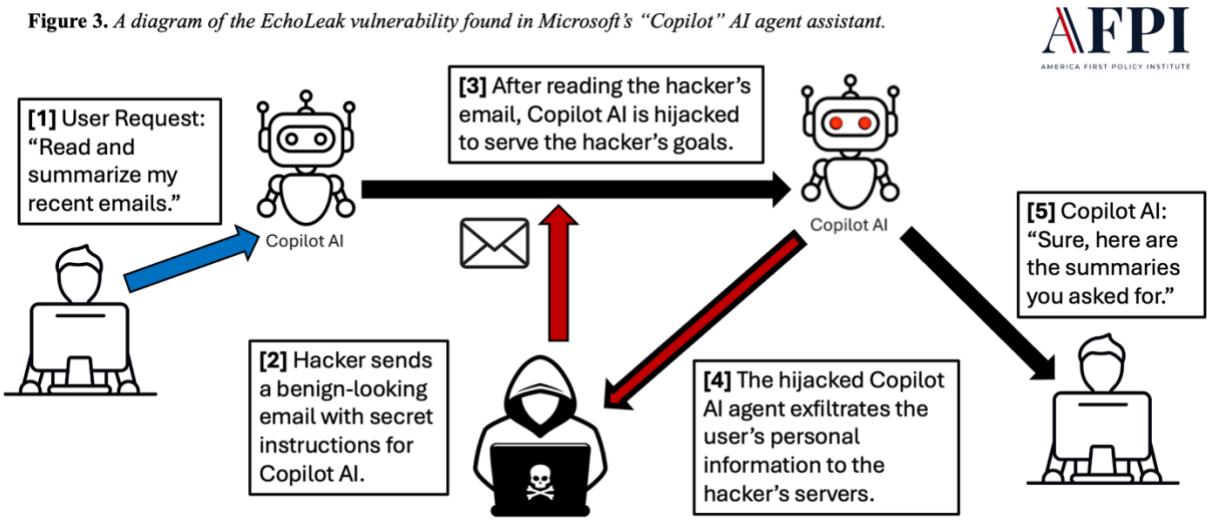
Agent Hijacking occurs when a malicious actor manipulates a deployed AI agent's behavior by embedding adversarial instructions within content the agent is directed to process. Attackers exploit AI agents' inability to reliably distinguish between trusted user instructions and untrusted data from their environment. The most common mechanism for agent hijacking is **indirect prompt injection**: the adversary plants a malicious prompt inside an innocuous-looking email, webpage, or document, which then overrides the original user's instructions and hijacks the agent to serve the adversary (Greshake et al., 2024). Researchers have found that all frontier models can be hijacked by these attacks (Chhabra et al., 2026).

Though non-agentic AI models can be hijacked, AI agent hijacking is far more concerning. For one, agentic models have a larger attack surface due to their web browsing tools and the multi-step, unpredictable nature of their tool use. Hijacking is also more consequential in an agentic context because agents can take consequential actions in real time. Research by the CAISI has demonstrated that hijacked agents can be induced to perform remote code execution on a user's machine, exfiltrate entire cloud file systems, and send personalized phishing emails to everyone in a user's contact list (CAISI, 2025a).

Fighting AI agent hijacking is a practical challenge at many enterprises today. One of the most prominent examples is the "EchoLeak" vulnerability of Microsoft's "Copilot" AI agent, discovered in mid-2025 (Reddy & Gujral, 2025). Copilot is an AI agent assistant with access to user data managed by Microsoft, such as Outlook emails, private files, and Teams chats, as well as tool-use capabilities such as sending emails and manipulating files. Copilot's integration with tools makes it useful but also leaves it vulnerable to indirect prompt injection.

In an "EchoLeak" attack, a malicious actor sends a legitimate-looking business email to a target's inbox. The email looks benign to human readers, but when Copilot reads its embedded text, the agent is hijacked to execute a complex exfiltration scheme to send private user data to attacker-controlled servers. Worryingly, the "EchoLeak" attack rendered traditional defenses like antivirus protections, firewalls, and static file scanning completely ineffective. The corrupting medium was pure text, and Copilot behaved exactly as it was programmed to: processing input and responding helpfully.

Figure 3. *A diagram of the EchoLeak vulnerability found in Microsoft's "Copilot" AI agent assistant.*



Note. The EchoLeak vulnerability is a real-world example of agent hijacking attacks, which co-opt an organization's internal AI agents to serve external malicious actors. Data from *EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System*, by Pavan Reddy and Aditya Sanjay Gujral, 2025 (https://arxiv.org/abs/2509.10540v1).

Today's defenses against agent hijacking are not robust. One recent experiment found that even with all existing defense techniques active, red teams can still extract sensitive information from AI systems using prompt injections (Abdelnabi et al., 2025). A growing concern is that hijacked agents can themselves become vectors for hijacking other AI systems in multi-agent environments, causing rapid cascades of security compromise that propagate before any security failure is detected (Gu et al., 2024; Motwani et al., 2025; Sharma et al., 2025; Hayum et al., 2026).

As AI agents become widely deployed in U.S. government systems, hijacking will become an ever-greater practical concern. For example, it will almost certainly prove useful for teams in the intelligence community to give AI agents access both to classified information and to open-source data scraped from the internet (Greshake, 2023). However, internet access also provides a clear mechanism by which such AI agents could be hijacked to exfiltrate classified data to malicious actors. Agent hijacking could be even more dangerous in contexts where AI agents are able to write and execute code in U.S. government codebases.

AI Deployment Security Protocols

Given the potential for AI agent hijacking, as well as undiscovered model poisoning, organizations that deploy AI agents should institute security protocols to prevent harmful actions. These protocols can take many different forms, but generally fall into the classes of (i) user oversight controls and (ii) system-level controls.

User Oversight Controls are a class of AI agent security strategies that rely on human operators to prevent or detect malfunctions or other anomalous actions. Generally, user oversight controls require human personnel to approve or verify sensitive agent actions in real time before they can occur. For example, some institutions require employees to manually approve all AI-generated code

before it can be executed (Wu et al., 2025). But user oversight controls have practical weaknesses that can render them unreliable when deployed alone.

User oversight controls can reduce the speed and usability of agents. Constant human review impedes AI agents, which are often favored because of their speed. This can result in automation bias, in which human users overly defer to agent requests rather than exercising independent judgment (Horowitz & Kahn, 2024). This challenge will increase as AI agents become more autonomous and complex. More complex oversight schemes will demand more of human overseers. This will place more practical pressure on them to defer to AI requests, defeating the purpose of oversight.

User oversight controls are also costly: employing human personnel to monitor AI agents and manually approve their actions can be prohibitively expensive. This oversight tax will increase with the number of agents an enterprise is using, the complexity of their workflows, and the granularity of the oversight scheme. Such costs can discourage AI agent adoption or push enterprises to overwork human monitors. For these reasons, user oversight control protocols are an insufficient solution to the agentic security challenge. But they can be valuable as one node in a network of system-level controls.

System-Level Controls are technical methods to prevent or detect malfunctions and other anomalous actions. Some researchers who predicted the rise of AI agents have explored methods that can improve the security of agents with minimal human oversight. Below, we provide a non-exhaustive list of system-level control techniques that have been supported by both research and real-world applications: (a) information access controls, (b) sandboxing, (c) audit logs, and (d) combined automated and human monitoring. These protocols have varying levels of adoption.

(a) Information access controls. Under this approach, users assume that any information available to an AI agent could leak. Therefore, AI agents are granted only the minimum level of data access necessary to complete their assigned tasks—a principle sometimes called "least privilege access" (Saltzer & Schroeder, 1975). This principle is commonly used to protect against human insider threats. If data is confidential, includes personal information, or is otherwise sensitive, deployers could place further guardrails before agents can access it. Information access controls are most effective when they operate within a larger institutional system of sensitivity labels.

(b) Sandboxing: hardware and software isolation between AI agents and data or execution modules (Wu et al., 2025). Without sandboxing, attackers could use hijacked AI agents to remotely execute arbitrary code—the ultimate threat to an organization's security (Banerjee & Aarne, 2026). For this reason, sandboxing is already a common practice in high-stakes enterprises. As Nvidia security staff advise: "Execution isolation is mandatory for AI-driven code execution. Sandboxing each execution instance limits the blast radius of malicious or unintended code. This control shifts security from reactive patching to proactive containment" (Irwin & Greshake, 2025).

(c) Audit logs of agent actions. Unlike standard user oversight controls, which require real-time human intervention, audit logging can operate asynchronously. The system records every action taken by an AI agent—tool calls, data accesses, external communications, file modifications—in a tamper-resistant log that can be reviewed after the fact. This approach allows security teams to reconstruct the precise sequence of events that led to any harmful or unauthorized AI agent action. It also empowers security teams to use those investigations to identify and patch systemic vulnerabilities in AI agent workflows, e.g., the vectors of a prompt injection attack. Audit logs should also include transcripts of AI reasoning, as they can be relevant to retrospective security investigations (Korbak et al., 2025).

(d) Automated agent monitoring. Researchers have explored how enterprises can build security protocols that use AI models to autonomously monitor other AI agents. A central finding in this field is that hybrid protocols—combining automated and human oversight—can allow institutions to reliably deploy AI models that may be untrustworthy or actively prone to misbehavior. Small, cheap, trusted AI models can effectively monitor the actions of larger, untrusted agents and flag certain behaviors for further review by automated or human systems. Researchers have even found that, using the right protocols, institutions can effectively use untrusted AI systems to monitor other untrusted AI systems (Greenblatt et al., 2024). Experiments show that certain strategies, such as resampling from untrusted agents, attempting to "honeypot" untrusted agents, and deploying filters to prevent agents from hijacking their automated monitors, can collectively enable institutions to effectively detect compromised AI agents and prevent them from causing harm—even with minimal real-time human oversight (Bhatt et al., 2025).

These deployment security protocols are not yet universally implemented in the private and public institutions that could benefit from the deployment of AI agents. This mismatch between the state of the research and the state of practice fuels concerns that AI agent insecurity is slowing adoption. The challenge is likely exacerbated in many government agencies, which are more risk-averse cultures than private enterprises. We recommend several policy actions to bolster agent security and accelerate adoption.

Solution: Promoting AI Deployment Security Standards

Solution 1: To accelerate AI agent adoption, CAISI could issue voluntary guidance on system-level AI agent security, both in low-stakes and high-stakes deployments. Such guidance should not be one-size-fits-all, but allow institutions to tailor their system-level AI agent security protocols to the individual vulnerabilities and needs of their specific AI workflows. For example, institutions that often handle personally identifiable information (PII) might focus on information access controls. CAISI could also ensure that its publications provide useful guidance to AI agent users in both high-stakes and low-stakes deployment environments. Some system-level AI agent security protocols that are appropriate in high-stakes deployment environments, such as U.S. national security agencies or frontier AI labs, are not necessary for commercial adopters of AI agent technology. CAISI's guidance on system-level AI agent security could encompass all the strategies discussed above, including information access controls, sandboxing, audit logs, and hybrid monitoring protocols. Voluntary guidance on securing AI agents would give enterprises and federal agencies the confidence they need to accelerate their deployment of the technology.

This guidance could also include open-source tools for system-level AI security protocols, such as automated monitoring. This would be a valuable service for government agencies and other institutions that lack the technical capacity to build their own agent security infrastructure.

Solution 2: CAISI could proactively encourage strong AI agent security standards in frontier AI labs. At many of the largest AI developers, internal AI agents now write the vast majority of new code (Nolan, 2026). In this context, sophisticated agentic security protocols are vital. An adversary who hijacks a frontier AI developer's AI agents could remotely sabotage the code that filters pre-training data, the systems that fine-tune AI models, the protocols for testing AI models for alignment and security, and many other essential safeguards in the AI development process (Banerjee & Aarne, 2026). Especially in partnership with compromised human insiders, hijacked internal AI agents at frontier AI labs could cause massive damage to AI development and American national security.

Unfortunately, frontier AI labs have inconsistently instituted system-level AI agent security protocols. One reason is that deploying agents with control methods can be expensive. As a Google engineer writes, "a key challenge for any monitoring system is to minimize a wide variety of additional costs, including computational resources, latency, privacy, availability, and complexity" (Shah et al., 2025). There is a coordination problem: even if each individual frontier AI company wants to institute high-quality AI agent security protocols, doing so alone may leave them with less money to invest in capabilities progress relative to their competitors. CAISI could work constructively with leading AI developers to promote the appropriate security standards for internal AI agents. End users of AI agents will be more confident in deploying the technology if they are assured that the AI developers are themselves working within strong agent security guidelines.

Solution 3: Congress could direct the Department of War to establish a program of record to develop cybersecurity measures for AI agents deployed in national security environments. The Pentagon is a leader in AI adoption, and is already moving to deploy AI agents at speed and scale. One of the core pillars of Secretary of War Pete Hegseth's AI strategy is "unleashing AI agent development and experimentation for AI-enabled battle management and decision support, from campaign planning to kill chain execution" (Hegseth, 2026). As this initiative progresses, proactively developing dedicated security measures will be essential to maintaining momentum. AI agents' autonomy, access to sensitive data, web browsing tools, and ability to interact with other agents may allow malicious actors to compromise American national security through agent hijacking. Dedicated cybersecurity measures and system-level agent security protocols can help the Pentagon deploy AI agent technology without fear of adversarial exploitation. Congress already directed the Department of War to "develop and implement a … policy for the cybersecurity and associated governance of artificial intelligence," including measures to combat "model serialization attacks, model tampering, data leakage, adversarial prompt injection, model extraction, model jailbreaks, and supply chain attacks" (S. 1071, 2025, Sec. 1512). A program of record would ensure that such policies continue to reflect the most effective security measures available to our warfighters.

Solution 4: Congress could direct the Department of Energy to develop and implement agent security protocols for use in the Genesis Mission. Launched by President Trump, the Genesis Mission is an initiative within the DOE to "train scientific foundation models and create AI agents to test new hypotheses, automate research workflows, and accelerate scientific breakthroughs." In service of this mandate, the DOE is constructing commercial-scale AI data centers and preparing vast proprietary government datasets—including classified datasets—to be deployed in AI agent workflows (DOE, 2025). The Genesis Mission will therefore be one of the first major federal initiatives to deploy AI agents in sensitive data environments. AI agents, however, introduce a novel attack surface. Top-secret systems today derive much of their security from features like air-gapping and compartmentalized access controls designed around human users. AI agents challenge both dimensions: they create new vectors for external exploitation and undermine internal controls that assume a human is in the loop. As discussed above, agents are also vulnerable to agent hijacking, which can allow malicious actors to co-opt them to actively leak proprietary information (CAISI, 2025a). In order to maintain the integrity of federal secrets and fulfill the mandate of the Genesis Mission, the DOE should institute strong security standards for AI agent deployment, including information access controls, sandboxing, audit logs, and hybrid human/automated monitoring protocols.

Risk: AI Agent Misbehaviors

Even if we prevent our adversaries from attacking American AI, AI agents can still misbehave due to technical malfunctions. For example, the AI agent that tried to "speedrun" deleting a Meta employee's email inbox was not hijacked by a malicious actor (Cramer, 2026). It was suffering from a technical failure that caused it to misunderstand or disregard her instructions.

Misbehaviors happen in part because the models powering AI agents are inherently nondeterministic and difficult to interpret (Olvera, 2025). This is not a problem if one is using a chatbot to generate poems, but it is a major security challenge if one is using an agentic system with tool-use capabilities and access to sensitive information. Unfortunately, AI developers lack reliable technical methods to make their models act in predictable and secure fashions. Developers have reduced the frequency of anomalous outputs and unpredictable actions, particularly through post-training. But developers do not yet have technical methods to produce models with the level of reliability that the U.S. government generally expects of weapons systems and critical infrastructure (Rosenblatt & Berg, 2025; Withers et al., 2026).

Solutions: Supporting AI Reliability Research

The research field of **AI reliability** aims to develop technical methods that allow AI developers to produce models that are less likely to malfunction and misbehave, more aligned with American interests, and more robust to adversarial attacks such as agent hijacking. It also provides AI agent users with technical methods to monitor and control agents. One useful way to think about AI reliability is that it ensures AI systems obey the proper "chain of command." Just as the U.S. military would not admit a service member who could not be trusted to reliably follow commands and advance American interests, the U.S. government should not be expected to deploy AI agents that cannot be trusted to reliably follow instructions.

Breakthroughs in AI reliability are public goods that benefit all users and developers of AI agent technology, so research is undersupplied by the private sector. The U.S. government has a central role to play in correcting this market failure by driving technical research in AI reliability. R&D investments in this field will pay dividends in public confidence in AI agent security and accelerate adoption of the technology across the economy and government.

AI reliability consists of several complementary research fields. These fields were mentioned explicitly in the White House's *AI Action Plan*, which called for the Defense Advanced Research Projects Agency (DARPA) and CAISI to direct investments in research related to "AI interpretability, AI control systems, and adversarial robustness" (Kratsios et al., 2025). As the Trump Administration recognizes, it is particularly urgent that breakthroughs in these areas occur before AI agent reliability becomes critical to the operation of U.S. critical infrastructure and national security institutions. To accelerate research in AI reliability, we recommend the following policy actions.

Solution 1: Congress could appropriate dedicated funding for AI reliability research through DARPA and the National Science Foundation (NSF). DARPA in particular has a proven track record of catalyzing breakthrough research in emerging technology domains, including in high-reliability technology (DARPA, n.d.), and AI reliability is among the most consequential technical challenges of the coming decade (Azoulay et al., 2019). Dedicated funding, on the order of tens of millions of dollars annually, could support research programs focused on:

- **AI interpretability**: methods for understanding the technical mechanisms behind model behavior, which remain largely unexplainable.
- **AI control**: protocols for constraining and monitoring AI agent behavior in deployment, including detecting malfunctions or adversarial behavior.
- **AI adversarial robustness**: techniques for AI developers and users to harden AI systems against poisoning, hijacking, or other attacks.

There is precedent for these programs: The White House's *AI Action Plan* already instructed federal agencies to invest in these three areas of technical research, and DARPA's Explainable AI program (2017-2021) laid the groundwork for future interpretability research (Kratsios et al., 2025; Gunning et al., 2021). Congress could further support the Trump Administration's push for breakthroughs in AI reliability, enabling multi-year research campaigns that are insulated from future political turnover.

Solution 2: DARPA and NSF could launch prizes for breakthroughs in AI interpretability, control, and adversarial robustness. Michael Kratsios, the director of the White House's Office of Science and Technology Policy (OSTP), affirms that: "In a moment of strategic significance, we must be more creative in our use of public research and development money, and shape a funding environment that makes clear what our national priorities are. … Prizes, advance market commitments, and other novel funding mechanisms, like fast and flexible grants, can multiply the impact of government-funded research" (Kratsios, 2025). The existing grant-based system may be inadequate for the urgency of the issue. In concert with research grants, DARPA and NSF could launch a collection of prize competitions for the most substantial research breakthroughs in AI interpretability, control, and adversarial robustness (Fist et al., 2025; Pistillo, 2025; for a private-sector version, see Upadhyay & Barez, 2025). Prizes would allow the U.S. government to effectively push the AI research community toward breakthroughs in areas that would be most advantageous for American prosperity and national security.

Summary of Policy Recommendations

AI Development Security Standards

The Department of Justice and the Federal Trade Commission should issue guidance clarifying that AI developers may share threat intelligence and coordinate security protocols without violating federal antitrust law. The most straightforward mechanism for accomplishing this goal would be reversing the Biden Administration's 2024 rollback of the "Antitrust Guidelines for Collaborations among Competitors," and adding specific guidance about security cooperation in the AI development industry.

CAISI should publish voluntary security frameworks and standards for frontier AI developers. These guidelines should recommend risk thresholds at which upgraded security standards may become appropriate. The recommended security standards should include:

- Data hygiene;
- Infrastructure security, including both cybersecurity and hardware security;
- Personnel security;
- Model auditing; and
- System-level security protocols for internal AI agent use.

NSA should lead red-teaming exercises that aim to uncover security vulnerabilities in the AI development supply chain. These exercises should be fully voluntary, and executed with the close cooperation of AI labs and data center operators.

The Department of War should implement AI procurement conditions that require strong AI development security standards to prevent model poisoning.

Congress should authorize the U.S. government to impose minimum security standards on large AI developers. These security standards should apply regardless of the developer's status as a U.S. government contractor.

Congress should pass whistleblower protections for staff who report security vulnerabilities at large AI developers.

AI Deployment Security Standards

CAISI should publish security frameworks for the deployment of AI agents in high-stakes and low-stakes environments, to enable the adoption of the technology across the economy. These guidelines should include information access controls, sandboxing, audit logs, and protocols for automated and human monitoring. They could also include open-source tools for AI agent oversight.

CAISI should proactively encourage strong AI agent security standards within frontier AI labs.

Congress should authorize a program of record in the Department of War to develop cybersecurity standards for AI agents deployed in national security environments.

The Department of Energy should publish and implement AI agent deployment security standards for use in the Genesis Mission.

AI Reliability Research

Congress should appropriate dedicated funding for AI reliability research through DARPA and the NSF. Funding commitments of around $50 million annually would enable these agencies to invest in multi-year research campaigns to improve the security of AI technology and accelerate its deployment. Relevant research areas include:

- AI interpretability, which aims to decipher the internal mechanisms behind AI behavior;
- AI control, which aims to discover and test protocols to secure AI agents; and
- AI adversarial robustness, which aims to secure AI agents against model poisoning and hijacking attacks.

DARPA and NSF should launch prizes for breakthroughs in AI interpretability, control, and adversarial robustness.

Conclusion

AI agents will create opportunities for a transformative leap in capabilities for American enterprises, government agencies, and warfighters. But security concerns are stalling their adoption. The U.S. government can break this impasse by promoting robust security standards in the AI development supply chain, publishing clear guidance on the secure deployment of AI agents, and supporting critical research in AI reliability techniques.

By moving decisively to implement the recommendations outlined in this brief, the U.S. government can ensure that AI agent technology is deployed rapidly, securely, and at a scale that will sustain American competitiveness and national security for decades to come.

REFERENCES

Abdelnabi, S. et al. (2025, June). *LLMail-Inject: A Dataset from a Realistic Adaptive Prompt Injection Challenge*. https://arxiv.org/abs/2506.09956

Altchek, A. (2024, June). *Ex-OpenAI employee speaks out about why he was fired: 'I ruffled some feathers'*. https://www.businessinsider.com/former-openai-researcher-leopold-aschenbrenner-interview-firing-2024-6

Antone, E. & Chang, A. (2025, March) *Cisco Introduces the State of AI Security Report for 2025: Key Developments, Trends, and Predictions in AI Security*. https://blogs.cisco.com/ai/cisco-introduces-the-state-of-ai-security-report-for-2025

Aschenbrenner, L. (2024, June). *Situational Awareness*. https://situational-awareness.ai

Azoulay, P. et al. (2019). *Funding Breakthrough Research: Promises and Challenges of the "ARPA Model"*. https://www.journals.uchicago.edu/doi/10.1086/699933

Banerjee, D. & Aarne, O. (2026, January). *AI Integrity: Defending Against Backdoors and Secret Loyalties*. https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/699e3adb0a9f1d4fb4728b20/1771977435150/AI+Integrity.pdf

Bhatt, A. et al. (2025, April). *Ctrl-Z: Controlling AI Agents via Resampling*. https://arxiv.org/abs/2504.10374

Bricken, T. et al. (2025, July). *Building and evaluating alignment auditing agents*. https://alignment.anthropic.com/2025/automated-auditing/

Carlini, N. et al. (2024, May). *Poisoning Web-Scale Training Datasets is Practical*. https://arxiv.org/abs/2302.10149

Center for AI Standards and Innovation (CAISI). (2025a, January). *Technical Blog: Strengthening AI Agent Hijacking Evaluations*. https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

Center for AI Standards and Innovation (CAISI). (2025b, September). *CAISI Works with OpenAI and Anthropic to Promote Secure AI Innovation*. https://www.nist.gov/news-events/news/2025/09/caisi-works-openai-and-anthropic-promote-secure-ai-innovation

Chhabra, A. et al. (2026, April). *Agentic AI Security: Threats, Defenses, Evaluation, and Open Challenges*. https://arxiv.org/abs/2510.23883

CNBC. (2026, March). *Watch CNBC's full interview with Department of Defense Undersecretary Emil Michael*. https://www.cnbc.com/video/2026/03/12/watch-cnbcs-full-interview-with-department-of-defense-undersecretary-emil-michael.html

Cramer, J. (2026, February). *'This should terrify you': Meta Superintelligence safety director lost control of her AI agent—it deleted her emails*. Fast Company.

https://www.fastcompany.com/91497841/meta-superintelligence-lab-ai-safety-alignment-director-lost-control-of-agent-deleted-her-emails

Cybersecurity and Infrastructure Security Agency (CISA). (2023, October). *NSA and CISA Red and Blue Teams Share Top Ten Cybersecurity Misconfigurations*. https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-278a

Defense Advanced Research Projects Agency (DARPA). (n.d.). *HACMS: High-Assurance Cyber Military Systems*. https://www.darpa.mil/news/resources/case-studies/hacms

Department of Energy (DOE). (2025, October). *Energy Department Announces New Partnership with NVIDIA and Oracle to Build Largest DOE AI Supercomputer*. https://www.energy.gov/articles/energy-department-announces-new-partnership-nvidia-and-oracle-build-largest-doe-ai

Department of Justice. (2026, January). *Former Google Engineer Found Guilty of Economic Espionage and Theft of Confidential AI Technology*. https://www.justice.gov/opa/pr/former-google-engineer-found-guilty-economic-espionage-and-theft-confidential-ai-technology

Doshi, R. (2024, September). *China's New National Security Laws: Risks to American Companies and Conflicts of Interest*. https://www.hsgac.senate.gov/wp-content/uploads/Testimony-Doshi-2024-09-24.pdf

Draganov, A. et al. (2026, February). *Phantom Transfer: Data-level Defences are Insufficient Against Data Poisoning*. https://arxiv.org/abs/2602.04899

Felstead, N. (2026, March). *How Antitrust Can Promote AI Safety Collaborations*. Lawfare. https://www.lawfaremedia.org/article/how-antitrust-can-promote-ai-safety-collaborations

Field, H. (2024, May). *OpenAI sends internal memo releasing former employees from controversial exit agreements*. CNBC. https://www.cnbc.com/2024/05/24/openai-sends-internal-memo-releasing-former-employees-from-non-disparagement-agreements-sam-altman.html

Fist, T. et al. (2025, March 17). *An Action Plan for American Leadership in AI*. Institute for Progress. https://ifp.org/an-action-plan-for-american-leadership-in-ai/

Freedberg, S. (2023, January). *NSA red team will attack JWCC providers to test zero trust security*. https://breakingdefense.com/2023/01/nsa-red-team-will-attack-jwcc-providers-to-test-zero-trust-security/

Freifeld, K. & Seetharaman, D. (2026, March). *Pentagon designates Anthropic a supply chain risk*. Reuters. https://www.reuters.com/technology/pentagon-informed-anthropic-it-is-supply-chain-risk-official-says-2026-03-05/

Gent, E. (2025, December). *AI coding is now everywhere. But not everyone is convinced.* https://www.technologyreview.com/2025/12/15/1128352/rise-of-ai-coding-developers-2026/

Ghaffary, S. (2024, June). *OpenAI Employees Want Protections to Speak Out on 'Serious Risks' of AI*. Bloomberg. https://www.bloomberg.com/news/articles/2024-06-04/openai-employees-call-for-protections-to-speak-out-on-ai-risks

Greenblatt, R. et al. (2024, July). *AI Control: Improving Safety Despite Intentional Subversion*. https://arxiv.org/abs/2312.06942

Greshake, K. [@KGreshake]. (2023, November). *PSA: The US Military is actively testing and deploying LLMs to the battlefield. I think these systems are likely to be vulnerable to direct prompt injection by adversaries* [Image attached] [Post]. [X]. https://x.com/KGreshake/status/1725515560870433016 https://x.com/KGreshake/status/1725515566369099805

Greshake, K. et al. (2024, January). *Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*. https://dl.acm.org/doi/epdf/10.1145/3605764.3623985

Grunewald, E. & Gershovich, A. (2025, September). *Accelerating AI Data Center Security*. Institute for AI Policy and Strategy. https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/68c025315a42f85b8206c465/1757422897410/Accelerating+AI+Data+Center+Security.pdf

Gu, X. et al. (2024, February). *Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast*. https://arxiv.org/abs/2402.08567

Gunning, D. et al. (2021, December). *DARPA's explainable AI (XAI) program: A retrospective*. https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.61

Hamin, M. & Edelman, B. (2025, November). *Cheating on AI Agent Evaluations*. https://www.nist.gov/caisi/cheating-ai-agent-evaluations

Harris, J. & Harris, E. (2025, April). *America's Superintelligence Project*. https://superintelligence.gladstone.ai/

Hayum, B., Egan, J., & Withers, C. (2026, March). *Recommendations for Securing and Promoting AI Agents*. https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/Agent-Security-RFI.pdf

Heckman, J. (2026, March). *AI boosts efficiency for agencies, but trust and safety lead the way*. Federal News Network. https://federalnewsnetwork.com/artificial-intelligence/2026/03/ai-boosts-efficiency-for-agencies-but-trust-and-safety-lead-the-way/

Hegseth, P. (2026). *War Department Launches AI Acceleration Strategy to Secure American Military AI Dominance*. https://www.war.gov/News/Releases/Release/Article/4376420/war-department-launches-ai-acceleration-strategy-to-secure-american-military-ai/

Horowitz, M. & Kahn, L. (2024, June). *Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts*. https://academic.oup.com/isq/article/68/2/sqae020/7638566

Hubinger, E. et al. (2024, January). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. https://arxiv.org/abs/2401.05566

Irwin, J. & Greshake, K. (2025, November). *How Code Execution Drives Key Risks in Agentic AI Systems*. https://developer.nvidia.com/blog/how-code-execution-drives-key-risks-in-agentic-ai-systems/

Jensen, B. & Strohmeyer, M. (2025, July). *Agentic Warfare and the Future of Military Operations*. https://www.csis.org/analysis/rethinking-napoleonic-staff

Kohli, P. & Lue, T. (2025, December). *Google DeepMind supports U.S. Department of Energy on Genesis: a national mission to accelerate innovation and scientific discovery*. https://deepmind.google/blog/google-deepmind-supports-us-department-of-energy-on-genesis/

Korbak, T. et al. (2025, July). *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*. https://arxiv.org/abs/2507.11473

Kratsios, M. (2025, April 14). *The Golden Age of American Innovation: Remarks by Director Kratsios at the Endless Frontiers Retreat*. The White House. https://www.whitehouse.gov/releases/2025/04/remarks-by-director-kratsios-at-the-endless-frontiers-retreat/

Kratsios, M., Sacks, D., & Rubio, M. (2025, July). *America's AI Action Plan*. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

Mahmood, Y., Salvador, C., & Crovitz, J. (2026, April). *China's Campaign to Steal and Sabotage American AI*. https://www.americafirstpolicy.com/issues/chinas-campaign-to-steal-and-sabotage-american-ai

Marks, S. et al. (2025, March). *Auditing language models for hidden objectives*. https://arxiv.org/abs/2503.10965

Michaels, S., Egan, J., & Daniel, M. (2026, March). *America's AI Cyber Defense Gap Needs Congress to Act*. https://www.cnas.org/publications/commentary/insights-americas-ai-cyber-defense-gap-needs-congress-to-act

Miyazono, E. (2025, August). *Preventing AI Sleeper Agents*. https://ifp.org/preventing-ai-sleeper-agents/

Motwani, S. et al. (2025, July). *Secret Collusion among AI Agents: Multi-Agent Deception via Steganography*. https://arxiv.org/abs/2402.07510

National Institute of Standards and Technology. (2026, January). *CAISI Issues Request for Information About Securing AI Agent Systems*. https://www.nist.gov/news-events/news/2026/01/caisi-issues-request-information-about-securing-ai-agent-systems

Nevo, S. et al. (2024, May). *Securing AI Model Weights.* RAND. https://www.rand.org/pubs/research_reports/RRA2849-1.html

Nolan, B. (2025, July). *An AI-powered coding tool wiped out a software company's database, then apologized for a 'catastrophic failure on my part'*. Fortune. https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure/

Nolan, B. (2026, January). *Top engineers at Anthropic, OpenAI say AI now writes 100% of their code—with big implications for the future of software development jobs*. Fortune. https://fortune.com/2026/01/29/100-percent-of-code-at-anthropic-and-openai-is-now-ai-written-boris-cherny-roon/

Olvera, A. (2025, January). *Why nobody can see inside AI's black box*. https://thebulletin.org/2025/01/why-nobody-can-see-inside-ais-black-box/

OpenAI. (2024, October). *Influence and cyber operations: an update*. Threat Intelligence Reports. https://openai.com/global-affairs/an-update-on-disrupting-deceptive-uses-of-ai/

Paulson Institute. (2020, June). *The Paulson Institute's Think Tank MacroPolo Releases Global AI Talent Tracker*. https://www.paulsoninstitute.org/press_release/the-paulson-institutes-think-tank-macropolo-releases-global-ai-talent-tracker/

Paulson Institute. (2024, March). *The Global AI Talent Tracker 2.0.* Edward Conrad Roundup. https://www.edwardconard.com/macro-roundup/57-of-the-top-2-of-global-ai-talent-currently-works-in-the-united-states-of-the-top-20-of-ai-workers-in-the-us-37-are-from-the-us-and-38-are-from-china-macropolochina/

Paulson Institute. (2025). *The Global AI Talent Tracker*. https://archivemacropolo.org/interactive/digital-projects/the-global-ai-talent-tracker

Pistillo, M. (2025, June 10). *Accelerating AI Interpretability To Promote U.S. Technological Leadership*. Federation of American Scientists. https://fas.org/publication/accelerating-ai-interpretability/

Reddy, P. & Gujral, A. (2025, September). *EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System*. https://arxiv.org/abs/2509.10540v1

Reinauer, A. (2026, March). *CEI Comments on NIST's Request for Information Regarding Security Considerations for Artificial Intelligence Agents*. https://cei.org/regulatory_comments/cei-comments-on-nists-request-for-information-regarding-security-considerations-for-artificial-intelligence-agents/

Request for Information Regarding Security Considerations for Artificial Intelligence Agents, 91 Fed. Reg. 698 (January 8, 2026). https://www.govinfo.gov/content/pkg/FR-2026-01-08/pdf/2026-00206.pdf

Rosenblatt, J. & Berg, C. (2025, December). *Can the U.S. Trust AI With National Security?* The Wall Street Journal. https://www.wsj.com/opinion/can-the-u-s-trust-ai-with-national-security-b481ac43

S. 1071. National Defense Authorization Act for Fiscal Year 2026. 119th Congress. (2025). https://www.congress.gov/bill/119th-congress/senate-bill/1071/text

Saltzer, J. & Schroeder, M. (1975, September). *The protection of information in computer systems*. https://ieeexplore.ieee.org/document/1451869

Senate Committee on the Judiciary. (2025, May). *Grassley Introduces AI Whistleblower Protection Act*. https://www.judiciary.senate.gov/press/rep/releases/grassley-introduces-ai-whistleblower-protection-act

Shah, R. et al. (2025). *An Approach to Technical AGI Safety and Security*. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/evaluating-potential-cybersecurity-threats-of-advanced-ai/An_Approach_to_Technical_AGI_Safety_Apr_2025.pdf

Sharma, G. et al. (2025, July). *Towards Unifying Quantitative Security Benchmarking for Multi Agent Systems*. https://arxiv.org/abs/2507.21146

Sheehan, M. & Zhuang, S. (2025, December). *Have Top Chinese AI Researchers Stayed in the United States?* https://carnegieendowment.org/emissary/2025/12/china-ai-researchers-us-talent-pool

Souly, A. et al. (2025, October). *Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples*. https://arxiv.org/abs/2510.07192

Trump, D. (2025a). *Ai.gov*. https://www.ai.gov/

Trump, D. (2025b, November). *Launching the Genesis Mission*. https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/

Upadhyay, S. & Barez, F. (2025). *Why Interpretability Matters.* Martian Interpretability Challenge. https://withmartian.com/prize

Vassilev, A. et al. (2025, March). *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. NIST. https://csrc.nist.gov/pubs/ai/100/2/e2025/final

Verma, P., Zakrzewski, C., & Tiku, N. (2024, July). *OpenAI illegally barred staff from airing safety risks, whistleblowers say*. https://www.stripes.com/theaters/us/2024-07-13/openai-prohibited-staff-safety-risks-14475707.html

Villalobos, P. et al. (2022, November). *Will we run out of ML data? Evidence from projecting dataset size trends*. https://epoch.ai/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset

Withers, C., Kim, J., & Chiu, E. (2026, March). *Off Target: A Working Paper on AI Alignment Challenges for National Security*. https://www.cnas.org/publications/reports/off-target

Wu, Y. et al. (2025, March). *IsolateGPT: An Execution Isolation Architecture for LLM-Based Agentic Systems*. https://arxiv.org/abs/2403.04960